平成 18 年度　　修　士　論　文　要　旨
教育研究分野　情報数理学　　6317028　福田 貴之

質的変数が利用可能なベクトル量子化法に基づく判別分析

1　序論

与えられたデータに対して，それがいずれのクラスに属するのかをより正しく判定する判別問題は従来より多変量解析の重要なテーマとして研究されてきた．これらはニューラルネットワークの学習機能や，ファジィシステムによるルールベースなどで判別を行っているが，多変量データが大きくなりすぎると判別能力に限界を生じさせてしまう欠点がある．

本論文では，判別を行う前段階として多変量データに対して効率的な正規化を行い，ベクトル量子化法の 1 つである LBG アルゴリズムを用いて判別を行う手法を提案する．この手法は多変量データの大きさにかかわらず判別が可能である．更に名義尺度や順序尺度で測定されるような質的変数を扱うための手法である数量化理論を適用する方法を提案し検証を行った．

2　LBG アルゴリズム

LBG アルゴリズムは Linde らが考案したクラスタリング手法である [1]．図 1 にクラスタリングの例を示す．ベクトル量子化によく用いられるアルゴリズムであり，具体的には以下の処理を行う．

1. データの中から要素を N 個選び初期代表ベクトルとする．
2. 全ての要素に対して，最も近い代表ベクトルを求め，クラスタリングを行う．
3. 同じクラスタに属する要素の重心を求め，新たな代表ベクトルとして更新する．
4. 代表ベクトルが更新されなければ終了，そうでない場合は 2. に戻る．

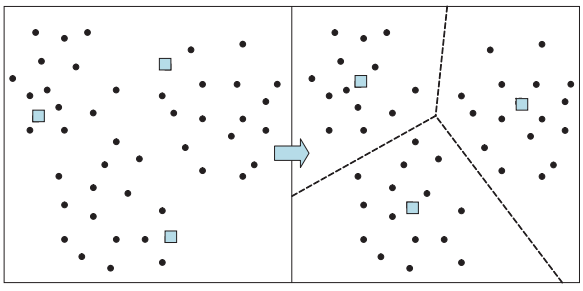

図 1: クラスタリング例

3　提案手法

LBG アルゴリズムを用いた判別手法の手順を以下に示す [2].

1. 与えられた多変量データに対して主成分解析を行い，有用な項目を n 個選択する．
2. 選択した項目 $q_1, q_2, \cdots, q_n$ を正規化する．
3. 正規化したデータ群の中の $q_{l1}, q_{l2}, \cdots, q_{ln}$ を 1 つのベクトルとみなし，LBG アルゴリズムによるクラスタリングを行う．
4. クラスタリング終了時，全ての要素がにどの代表ベクトルに属しているかを調べることで判別を行う．

本論文第 3 章では提案した判別システムそれぞれについて具体的な説明を行う．更に質的変数の組み込みについて説明する．

3.1　項目の選択手法

与えられた多変量データが有している情報をより解釈しやすくするために主成分分析を行う．主成分解析の流れは以下の通りである．

1. データが持つ成分間の相関を示す相関係数を求め，それを行列で表した相関係数行列 R を求める．
2. 相関係数行列 R の第 1 固有値（最大固有値）λ_1 に対応する固有ベクトルから第 1 主成分 z_1 を求める．同様にして，R の第 i 固有値 $\lambda_i (i = 1, 2, \cdots, n)$ に対応する固有ベクトルから第 i 主成分 z_i を求める．
3. 1 つの主成分が元の全変数が持っている情報の何割を説明できるかを表す指標である寄与率を求める．

$$\text{第 } i \text{ 主成分の寄与率} = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \tag{1}$$

以上により求めた寄与率により判別に用いる成分を決定する．

3.2　データの正規化

3.1 で選択したデータに対して正規化を行う．本論文では，データの中に外れ値のような値があってもその特徴をうまく正規化できる手法である 3 シグマ法による変換を用いた．

データを $q = \{q_1, \cdots, q_n\}$，変換後の出力 $q^* = \{q_1^*, \cdots, q_n^*\}$，また平均を $E(q)$，分散を $Var(q)$ とし，求める変換の式を

$$q^* = aq + b \tag{2}$$

とする．仮に変換後のデータを 0 から 255 までの値とし，平均を 128 とすると次式が求まる．

$$\begin{aligned} E(aq+b) &= a\bar{q} + b \\ &= 128 \end{aligned} \tag{3}$$

次に分布の広がりであるが，正規分布の場合，平均 ±3 シグマに全体の 99.7%が入る．本データは正規分布に従うという保証はないが，それでも 3 シグマのレンジを考えれば，ほとんどのデータがこれに含まれると考えられる．したがって $Var(q^*)$ は式 (4) で求まる．

$$Var(aq+b) = a^2\sigma^2 \tag{4}$$

ここで 3 シグマ分のレンジを 120 とすれば，標準偏差 $\sigma = 40$ となり，式 (4) により式 (5) が求まる．

$$a\sqrt{Var(q)} = 40 \tag{5}$$

式 (3),(5) より，a, b を求めることで変換式が一意に求まる．

3.3 質的変数の組み込み

数値で表される量的データのみで構成された多変量データを用いて判別を行っているが，それだけでは判別が困難な場合，問診やアンケートといった質的データを収集することで判別の補助が出来ると考えられる．

いま，表 1 のような A または B グループに属している人に対してアンケート形式で質問を行いデータを収集したとする．

表 1: アンケート結果例 (一部)

No.	Q1	Q2	グループ
1	Y	Y	A
⋮	⋮	⋮	⋮
6	Y	N	B

各アイテムの各カテゴリについて，ダミー変数として，Y と回答は 1，N と回答は 0 の数量を与える．これらのカテゴリから求められる総合量を y とすると，

$$y = a_0 + a_1x_1 + a_2x_2 \tag{6}$$

となり，相関比を最大とするような a_i を求めれば，回帰式が一意に求まる．

予め回帰式を求めておき，新たに収集した質的データを量的データとして変換する．そのデータを今ある多変量データと同様に 0〜255 の値に正規化を行い項目の 1 つとして判別に用いた．

4 検証

4.1 ワイン判別問題への適用

ワインの判別問題は 3 銘柄の予め与えられている 13 次元からなるワインの成分データを用いてどのワインであるか判別する問題である．先行研究としてファジークラスタリングによる判別などが行われている [3]．ワインデータの各クラスは 59，71，48 個，総計 178 個のデータからなる．正しく代表ベクトルに属しているかどうかを判別率として計算しその結果を表 2 に示す．

表 2: ワイン判別結果

種類	識別率（%）
ワイン 1	100.0
ワイン 2	85.7
ワイン 3	90.5

提案手法による判別の結果，全体で 92.1%の判別率であった．これは従来法 [3] の判別率が 97.2%であることを考えると若干劣るものの，提案手法がワイン判別に対してある程度効果があると確認できる．

4.2 肝臓病診断問題への適用

肝臓病判別問題は肝臓病の 5 段階の病状，健康体→急性肝炎→慢性肝炎→肝硬変→肝臓癌のどの段階であるか判別する問題である [4]．患者は 18 項目の血液検査結果が与えられている．まず血液検査データのみで健康体，急性肝炎，肝硬変の 3 種類のクラスを判別する．

更に，患者に対して 14 項目からなる問診を実施し，その問診データを組み込み 5 種類全てのクラスを判別を試みた．問診結果を表 3 に示す．また判別結果を表 4，5 に示す．

表 3: 問診結果（一部）

No.	Q1	Q2	···	Q14	グループ
1	Y	N	···	N	健康体
2	N	N	···	N	健康体
⋮	⋮	⋮	···	⋮	⋮
256	Y	Y	···	Y	肝臓病

表 4: 3 種類判別結果

種類	識別率 (%)
健康体	97.7
急性肝炎	89.4
肝硬変	71.7

表 5: 5 種類判別結果

種類	識別率 (%)
健康体	100
急性肝炎	66.6
慢性肝炎	74.5
肝硬変	82.4
肝臓癌	92.2

質的変数を組み込まない 3 種類判別の結果，全体で 86.3%であった．従来法であるアソシアトロンによる 3 種類判別の判別率が 73.3%であることを考えると，全体的に提案法が有効であることが確認できる．

また，質的変数を組み込んだ 5 種類判別の結果であるが，架空の問診結果ではあるが，形式的なチェック項目による問診を組み込むことで，判別できなかった病種を高い判別率で判別できていることがわかる．

5 結論

本論文では質的変数が組み込み可能なベクトル量子化法に基づく判別方法を提案し，実際の検証例として，ワインの判別問題と肝臓病の診断問題に適用した．

今後の課題として，肝臓病診断問題では問診結果を Y/N の 2 値の質的データで収集を行っていたが，3 値以上のより拡張した回答方式が望まれる．また実際の例に基づく検証を行う必要がある．

参考文献

[1] Y.Linde,A.Buzo and R.M.Gray,“An algorithm for vector quantizer design”,IEEE Trans. Commun.,Vol.COM-28,No.1,pp.84-95(1980)

[2] 福田貴之, 山内 仁, 金川明弘, 高橋浩光,“LBG アルゴリズムを用いた判別問題の一解法”, 第 8 回 IEEE HISS 論文集,pp.200-201(2006)

[3] 山本康高, 吉川大弘, 古橋 武,“判別分析を基準とするファジークラスタリングによる多次元データの可視化手法の提案”, 信学論,Vol.J88-D-II,No.6,pp.975-984(2005)

[4] A.Kanagawa,H.Kawabata,H.Takahashi,“Cellular neural networks with multiple-valued output and its application”, IEICE Trans. on Fundamentals,Vol.E79-A,No.10,pp.1658-1663(1996)